DOCUMENT RESUME

ED 139 820 TM 006 309

AUTHOR Fishbein, Ronald L.
TITLE The Fairness of Test Items for Statistically Equated Groups.
PUB DATE [Apr 77]
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement; *Analysis of Covariance; Basic Skills; *Black Students; *Caucasian Students; Grade 8; *Matched Groups; Racial Differences; *Test Bias

ABSTRACT

A nonstandard analysis of covariance (homogeneity of regression not assumed) was performed on 23 black-white item comparisons that had manifested bias according to an item X group interaction definition of bias, and on 9 black-white comparisons that had not manifested bias, to determine whether item bias is reduced for white and black examinees whose achievement levels are equated statistically. Results indicated that test items are less biased for these examinees if their achievement levels are equated statistically. The subjects for this study were black and white examinees who participated in the 1973 Florida Eighth-Grade Testing Program. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED139820

The Fairness of Test Items for Statistically Equated Groups*

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Ronald L. Fishbein
United States Office
of Education**
Office of Planning, Budgeting
and Evaluation

Abstract

This study examined whether test items are less biased for White and Black examinees whose achievement levels are equated statistically than for the same examinees if their achievement levels are not equated statistically. The results indicated that, in general, test items are less biased for Black and White groups whose achievement levels are equated statistically than for the same examinees if their achievement levels are not equated statistically. Furthermore, the results suggested that as the difference in the overall performance between Blacks and Whites in a subject area is reduced, so may the degree of bias on tests in that subject area.

*Paper presented at the 1977 meeting of the American Educational Research Association, New York, New York, April 1977 (Session 16.14)

**This paper was written by Ronald L. Fishbein in his private capacity. No official support or endorsement by the U.S. Office of Education is intended or should be inferred.

TM006 309

This study attempted to determine whether reading, mathematics, and occupational information test items are less biased for White and Black examinees whose achievement levels are equated statistically than for the same examinees if their achievement levels are not equated statistically. That is, a test item was considered biased if, after attempting to equate statistically the achievement levels of the two groups being compared, there was a statistically significant difference in the predicted performance of the two groups on the item. The degree to which one can equate this operational definition of bias (fairness) with "cultural bias" (fairness) is directly related to the degree that one can accept the assumption that two groups of equal overall ability in a subject area should not perform significantly different on a test item in that subject area.

Angoff and Ford (1973) attempted to extend the Item X Group interaction method of assessing the fairness of a test to consider the relationship between test bias and the difference in test performance between Black and White examinees. The primary question which Angoff and Ford's study attempted to answer was whether matching samples across ethinc groups, on some ability related to the ability required on the test being assessed for bias, would reduce the Item X Group interaction (the measure of bias) between two ethnic groups. Although no statistical tests were employed, their findings suggested that some of the interaction on both the verbal and mathematical sections of the Preliminary Scholastic Aptitude Test was related to the difference in test performance between the Black and White examinees studied. That is, correlations between the scores of Black and White groups matched on achievement level

were higher than the correlations between Black and White groups whose achievement levels had not been matched. This indicated the need for further research to determine whether Angoff and Ford's preliminary findings could be duplicated in a study using more rigorous statistical procedures, and suggested the definition of bias employed in this study -- groups of equal overall ability in a subject area should not perform significantly different on an item in that subject area.

Method

Samples

Samples were chosen from the data of the 1973 administration of the Florida Eighth Grade Testing Program (FEGTP), a basic skills test battery administered annually to virtually every eight grade student in Florida. Five systematic samples of 225 examinees in each sample were chosen from the population of White examinees and five systematic samples of 225 examinees in each sample were chosen from the population of Black examinees. Examinee classifications were determined from the information provided by the examinee on the answer sheet for the test battery under the category Race Code. All samples in both populations were chosen without replacement. Since the item responses of the examinees who participated in the testing program were grouped by county, and by school within each county, the systematic samples chosen were, for practical purposes, equivalent to stratified samples where the number of examinees chosen from any school represented approximately the proportion of examinees from that school in the population of the group sampled.

Procedure

A nonstandard analysis of covariance (ANCOVA) was performed on 23 Black-White item comparisons that had manifested bias according to an Item X Group interaction procedure used in an earlier study by the author (see Fishbein, 1975). The ANCOVA procedure was also used to assess 9 Black-White item comparisons that had not manifested bias.

The following mathematical model was employed, where μ is the adjusted grand mean of the treatment populations; τ_i is the effect

$$Y_{ij} = \mu + \tau_i + \beta_i + (X_{ij} - \overline{X}) + \varepsilon_{ij}$$

of the ith treatment on the dependent variable; β_i is the slope of the linear regression of the dependent variable Y on the covariable X for the ith group; X_{ij} is the value of the covariable for the jth individual of the ith treatment group; $\overline{X}$ is the overall mean of the covariable; and ε_{ij} is the error in the score of the jth individual of the ith group. The model is considered nonstandard because it does not make the usual assumption of homogeneity of regression ($\beta_1 = \beta_2 = \ldots = \beta$).

Specifically, the researcher let $\mu_{y(vim):x = qr}$ correspond to the mean value of the dependent variable at a covariable score of qr for the vth test for the ith group on the mth item. Then, to test the equality of the population means, confidence intervals were constructed. If a confidence interval did not span zero, the null hypothesis was rejected and the item was considered biased against the group with the lower mean value on the dependent variable at the covariable score of interest.

In this study, 32 items were tested for bias at three points (high, medium, and low) on a score continuum of the covariable that was derived by using linearly estimated true scores (see Nunnally, 1967). The α level for each test was .001. The low, medium, and high points were determined from a frequency count of the estimated true scores obtained by the Black and White examinees of the first Black and White samples on the test that was used as the covariable. The test that was used as the covariable was the test in the battery most highly correlated with the test that was assessed for bias. Because few Black examinees sampled obtained high scores on the tests used as covariables, the high point was often below the White group's mean on the covariable.

Results

A nonstandard ANCOVA was performed on 23 Black-White comparisons which had manifested bias according to an Item X Group interaction definition of bias and on 9 Black-White comparisons which had not manifested bias. For the 15 Black-White comparisons which had previously exhibited bias against Black examinees, the ANCOVA procedure failed to detect a significant difference between the population means at the three levels tested for six of the items; significant differences were detected at the middle and upper levels, but not at the lower level, for six of of the items; a significant difference was detected at the lower level, but not at the middle or upper levels, for two of the items; and significant differences were detected at all three levels for one of the items. In all such instances, the ANCOVA procedure revealed bias against Blacks.

An explanation of why the majority of instances where the null hypothesis was rejected occurred at the medium and high scores on the covariable continuum was not apparent.[1] The null hypothesis was not rejected for any of the eight items previously biased against Whites.

The results of the ANCOVA tests on the nine items which had failed to show interaction indicated that seven of the items failed to reject the null hypothesis at the three levels tested. The null hypothesis was rejected on one test item at the low and middle levels and, on a second item, at the high level of the covariable. In these three cases, the comparisons were biased against the Black group.[2]

Discussion

The ANCOVA procedure was used to remove statistically whatever differences between two groups on a dependent variable could be predicted from the difference between the two groups on an independent variable (covariable) associated with the dependent variable. One of the requirements of a standard or nonstandard fixed effects ANCOVA is that the subjects have been randomly assigned to the experimental groups, and the groups differ only in receiving different treatments. If this requirement is violated, as it must be when working with ethnic groups, then the ANCOVA technique probably is unable to adjust completely for preexisting differences between two groups. This is especially true when differences are large, such as between White and Black examinees. Cochran (1957) has stated that although ANCOVA can remove bias from an experiment which has employed naturally occurring groups, one can never be sure that all preexisting differences between two groups have been removed. Although

an ANCOVA adjustment with randomization does not guarantee that the groups have been equated, the effects of a biased variable have probably been nearly equally distributed among the treatment groups. Stated simply, testing whether an item operates differentially on Black and White examinees who have the same obtained score on a variable highly related to the test being examined for bias, is not the same as equating the two groups on all factors which affect their performance on an item--or, as stated previously, having equal overall ability in a subject area. Perhaps a principal axis adjustment (see Wainer, 1976), which seems to be a sounder procedure than ANCOVA for statistically adjusting differences between nonrandom groups, would have been more appropriate. Unfortunately, the researcher was not familiar with the principal axis adjustment when this study was conducted.

Conclusions and Recommendations

The results of this study indicated that, in general, test items are less biased for White and Black examinees whose achievement levels in a subject area are equated statistically than for the same examinees if their achievement levels are not equated statistically. These results, which are consistent with Angoff and Ford's, suggest that the difference in performance levels between White and Black examinees may be related to the degree of bias shown on educational tests. That is, as the difference in achievement levels between Blacks and Whites is reduced, the degree of bias on educational tests may be reduced also. This seems to indicate that the long-term resolution of the test bias issue may lie more in upgrading the educational opportunities of minority-group children than

in the development of statistical techniques to detect biased items. When minority-group is no longer synonomous with lower-scoring group, the issue of test bias may cease to exist. Nevertheless, the results of this study do have practical implications for the test developer.

Implications for the test constructor

The tests used in the FEGTP were professionally developed, had undergone considerable editing, and had been administered to representative Florida samples before the final versions were printed. Even with these safeguards, the Item X Group interaction procedure used by Fishbein (1975) detected some biased group comparisons; and some of those biased comparisons examined in this study (with a much less stringent definition of bias) remained biased. In the opinion of this researcher, it would have been extremely advantageous if biased items could have been detected during the test development stage.

It is recommended that a statistical procedure be employed during the test development stage to detect items which may be biased against certain ethinc groups. An Item X Group interaction procedure might be applied, and those items detected as biased might then be tested with the procedure used in this study or a principal axis adjustment. Items detected as biased by both procedures could be examined and, if flaws were found, they could, hopefully, be corrected; such items could be reanalyzed with a new sample of subjects until a satisfactory result was reached. However, it is possible that items detected as biased, by both an Item X Group interaction definition and by a method which statistically equates ethnic

groups on achievement, may not be biased and should not be revised or eliminated from a test. Statistical tests for bias may be useful screening devices but should not be used arbitrarily.

Footnotes

1) It should be noted that when employing nonstandard ANCOVA, the point on the covariable where the standard error of the difference between the two predicted means is smallest is between the means of the two groups. The standard error was probably smaller at the medium point on the covariable than at the high point or the low point. However, plots of regression lines were generally consistent with the significant tests.

2) However, it should be noted that the α levels for the Item X Group interaction tests were .0002 compared to .001 for the ANCOVA tests. Thus, if .0002 α levels were used for the ANCOVA comparisons, the null hypothesis may not have been rejected for some of the items which failed to show bias with the interaction technique.

References

Angoff, W.H., & Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Cochran, W.G. Analysis of covariance: Its nature and use. Biometrics, 1975, 13, 261-281.

Fishbein, R.L. An investigation of the fairness of the items of a test battery. Paper presented at the meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83 (2), 213-217.